Inference and Variable Selection for Random Forests

Lucas Mentch
Department of Statistics
University of Pittsburgh

Banff International Research Station
Workshop on the Interface of Machine Learning and Statistical Inference

January 2018

Collaborators

- Cornell University, Department of Statistics
  - Giles Hooker
  - Graduate Students: Sarah Tan, Yichen Zhou
- Cornell University, Lab of Ornithology
  - Dan Fink, Frank La Sorte, David Winkler, Wesley Hovhachka
- University of Pittsburgh, Department of Statistics
  - Graduate Students: Tim Coleman, Wei Peng

Outline

1. Background & Motivation
   - What we're trying to do and why
2. Asymptotic distributional results
   - Central Limit Theorems for Subsampled Ensembles; Confidence Intervals for predictions
3. Hypothesis testing
   - Tests for feature importance/significance
   - Tests for model additivity/interactions
4. Variable Selection & Importance Measures
   - Hold-out forests
5. Ebird Application

Background & Motivation

General Supervised Learning Set-up

Notation:

- Response Y
- Features (covariates) $\mathbf{X} = \{X_1, \ldots, X_p\}$
- Prediction point (feature vector) $\boldsymbol{x}^*$
- Prediction $\hat{y}^* = \hat{F}(\boldsymbol{x}^*) \in \mathbb{R}$

- We assume we have an i.i.d. training set

$$T_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$$

used to construct the prediction function $\hat{F}$, where

$$Y_i = F(\mathbf{X}_i) + \epsilon_i, \qquad \epsilon_i \overset{iid}{\sim} mean\ 0$$

Machine Learning Pros & Cons

- Given this generic situation with “a lot" of data, but limited *a priori* intuition with respect to underlying relationships in the data, ML tools present an attractive path forward:
  - Little to no model specification often required
  - Properly tuned models can produce *very* accurate predictions

  But ...

  - Computing $\hat{F}$ may be computationally expensive
  - Limited ability to do inference; loss of intuition (“Black-boxes")
  - Few if any theoretical gaurantees

Machine Learning Pros & Cons

- So what gets done in practice?
  - Reliance on (scarcely available) *ad hoc* tools
  - “Forced" (improper) application of classic statistical tests (“We **will** get a p-value one way or the other")
  - Use ML for predictions, simpler (usually linear) statistical models for inference

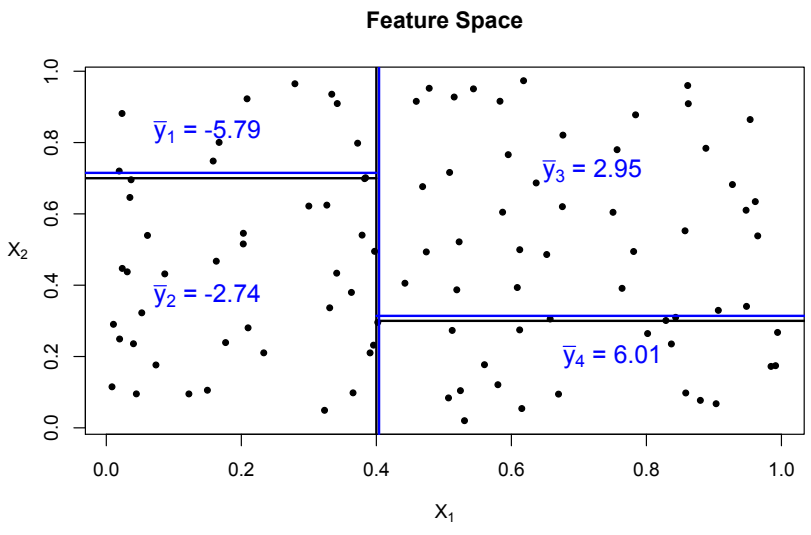
And thus, what we’d like is ...

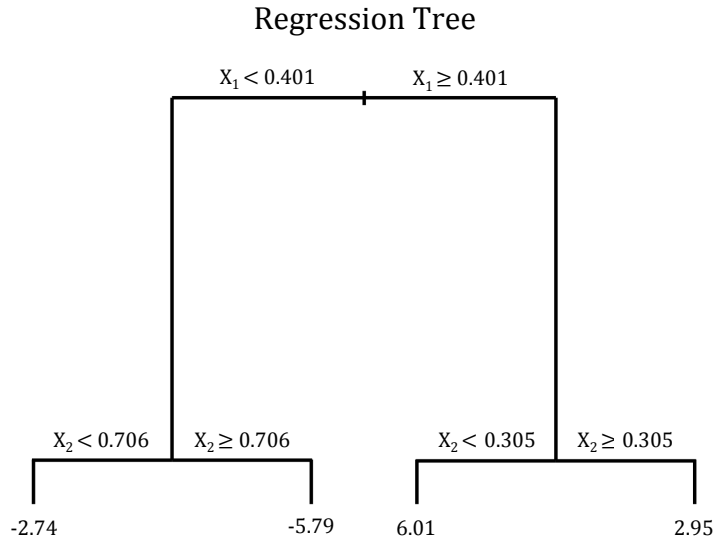
  - Computationally efficient set of inferential tools for better understanding underlying relationships in the data within traditional “black-box" contexts that come with some statistical and mathematical backing

Random Forests

Individual Trees

- Trees built by sequentially partitioning the feature space





- Splits chosen greedily to most improve predictions

  $\Longrightarrow$ High variance
- Tendency to over-fit; define cost-complexity parameter
- Very difficult to analyze.

Tree-Based Ensembles

- Ensembles of trees usually stabilize variance and improve predictions
  - Bagging (**b**ootstrap **agg**regat**ing**) - take B bootstrap samples of training set, build a tree with each new sample, and average over predictions from each tree to get final prediction

$$\hat{Y}^*_B = \frac{1}{B}\sum_{i=1}^{B} T_{\boldsymbol{x}^*}((\boldsymbol{X}_{i_1}, Y_{i_1}), \ldots, (\boldsymbol{X}_{i_n}, Y_{i_n}))$$

  - Random Forests - similar to bagging, but at each potential split point in each tree, select the best variable to split based on a random selection of only $d < p$ features.

$$\hat{Y}^*_{RF} = \frac{1}{B}\sum_{i=1}^{B} T_{\boldsymbol{x}^*,\omega_i}((\boldsymbol{X}_{i_1}, Y_{i_1}), \ldots, (\boldsymbol{X}_{i_n}, Y_{i_n}))$$

Random Forests

Why random forests?

- They work! Top ranked method across 100's of classifiers (Fernández-Delgado et al. 2014) *and* amongst the best "off-the-shelf"
- Nice macroscopic structure

Why are they so difficult to analyze?

1. Greediness in fitting makes obtaining distributional results for individual trees extremely difficult
    - Adding deterministic structure gets us back to traditional statistics
2. Bootstrapping compounds the correlation issues

Distributional Results and CIs

Subsampling

Idea: Construct trees with m_n *subsamples* of size k_n instead of full bootstrap samples and structure ensemble instead of base learners

$$\hat{F}(\boldsymbol{x}^*) = \frac{1}{m_n} \sum_{i=1}^{m_n} T_{\boldsymbol{x}^*}((\boldsymbol{X}, Y)_{i_1}, \ldots, (\boldsymbol{X}, Y)_{i_{k_n}})$$

- Looks a lot like a U-statistic, but need to extend results to (possibly randomized) kernels with growing rank
- **Trade-off:** Want subsamples to be big enough so that trees can grow large enough to capture sufficient signal, but small enough that dependence is manageable

CLT for Subbagged Ensembles

Theorem 1

Let $Z_1, Z_2, \ldots \overset{iid}{\sim} F_Z$ and let U_{n,k_n,m_n} be an incomplete, infinite order U-statistic with (Lipschitz) kernel h_{k_n}. Let $\theta_{k_n} = \mathbb{E}h_{k_n}(Z_1, \ldots, Z_{k_n})$ such that $\mathbb{E}h^2_{k_n}(Z_1, \ldots, Z_{k_n}) \leq C < \infty$ for all n and some constant C, and let $\lim \frac{n}{m_n} = \alpha$. Then as long as $\lim \frac{k_n}{\sqrt{n}} = 0$ and $\lim \zeta_{1,k_n} \neq 0$,

(i) if $\alpha = 0$, then $\frac{\sqrt{n}(U_{n,k_n,m_n} - \theta_{k_n})}{\sqrt{k_n^2 \zeta_{1,k_n}}} \xrightarrow{d} \mathcal{N}(0,1)$.

(ii) if $0 < \alpha < \infty$, then $\frac{\sqrt{m_n}(U_{n,k_n,m_n} - \theta_{k_n})}{\sqrt{\frac{k_n^2}{\alpha}\zeta_{1,k_n} + \zeta_{k_n,k_n}}} \xrightarrow{d} \mathcal{N}(0,1)$.

(iii) if $\alpha = \infty$, then $\frac{\sqrt{m_n}(U_{n,k_n,m_n} - \theta_{k_n})}{\sqrt{\zeta_{k_n,k_n}}} \xrightarrow{d} \mathcal{N}(0,1)$.

Conditions

Condition 1: Let $Z_1, Z_2, \ldots \overset{iid}{\sim} F_Z$ with $\theta_{k_n} = \mathbb{E}h_{k_n}(Z_1, \ldots, Z_{k_n})$ and define $h_{1,k_n}(z) = \mathbb{E}h_{k_n}(z, Z_2, \ldots, Z_{k_n}) - \theta_{k_n}$. Then for all $\delta > 0$,

$$\lim_{n\to\infty} \frac{1}{\zeta_{1,k_n}} \int_{|h_{1,k_n}(Z_1)| \geq \delta\sqrt{n\zeta_{1,k_n}}} h_{1,k_n}^2(Z_1) dP = 0.$$

Proposition 1: For a bounded regression function F, if there exists a constant c such that for all $k_n \geq 1$,

$$\left|h((\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_{k_n+1}, Y_{k_n+1})) - h((\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_{k_n}, Y_{k_n}), (\boldsymbol{X}_{k_n+1}, Y^*_{k_n+1}))\right| \leq c\left|Y_{k_n+1} - Y^*_{k_n+1}\right|$$

where $Y_{k_n+1} = F(\boldsymbol{X}_{k_n+1}) + \epsilon_{k_n+1}$, $Y^*_{k_n+1} = F(\boldsymbol{X}_{k_n+1}) + \epsilon^*_{k_n+1}$, and where ϵ_{k_n+1} and $\epsilon^*_{k_n+1}$ are i.i.d. with exponential tails, then Condition 1 is satisfied.

CLT for Random Forests

Theorem 2

Let $U_{\omega;n,k_n,m_n}$ be a random kernel U-statistic such that $U^*_{\omega;n,k_n,m_n}$ satisfies Condition 1 and suppose that $\mathbb{E}h^2_{k_n}(Z_1,\ldots,Z_{k_n}) < \infty$ for all n, $\lim \frac{k_n}{\sqrt{n}} = 0$, and $\lim \frac{n}{m_n} = \alpha$. Then, letting β index the subsamples, so long as $\lim \zeta_{1,k_n} \neq 0$ and

$$\lim_{n\to\infty} \mathbb{E}\left(h^{(\omega)}_{k_n}(Z_{\beta_1},\ldots,Z_{\beta_{k_n}}) - \mathbb{E}_\omega h^{(\omega)}_{k_n}(Z_{\beta_1},\ldots,Z_{\beta_{k_n}})\right)^2 \neq \infty,$$

$U_{\omega;n,k_n,m_n}$ is asymptotically normal and the limiting distributions are the same as those provided in Theorem 1.

- We've got the distributions; once we estimate the parameters, we can pull out confidence (prediction) intervals. On to more exciting kinds of inference!

Hypothesis Testing

Testing Feature Significance

- Partition collection of features $\{X_1, \ldots, X_p\}$ into two sets:

$$\boldsymbol{X}_R := \textbf{R}\text{educed set of features known to be important}$$
$$\boldsymbol{X}_A := \textbf{A}\text{dditional features to test for importance}$$

in order to test

$$H_0 : F(\boldsymbol{X}_R, \boldsymbol{X}_A) = F_R(\boldsymbol{X}_R) \text{ at } N \text{ total test points}$$

- Let $\hat{F}$ denote the original ensemble, and $\hat{F}_R$ denote the ensemble that ignores $\boldsymbol{X}_A$, so that we have

$$\begin{aligned}
\hat{D}(\boldsymbol{x}_i^*) &= \hat{F}(\boldsymbol{x}_i^*) - \hat{F}_R(\boldsymbol{x}_i^*) \\
&= \frac{1}{m_n} \sum_j T_{\boldsymbol{x}_i^*}(S_j) - \frac{1}{m_n} \sum_j T_{\boldsymbol{x}_i^*, R}(S_j) \\
&= \frac{1}{m_n} \sum_j \left(T_{\boldsymbol{x}_i^*}(S_j) - T_{\boldsymbol{x}_i^*, R}(S_j) \right)
\end{aligned}$$

- This is a U-statistic, so $\hat{\boldsymbol{D}}^T \hat{\Sigma}_D^{-1} \hat{\boldsymbol{D}} \sim \chi^2_N$ can be used as a test statistic.

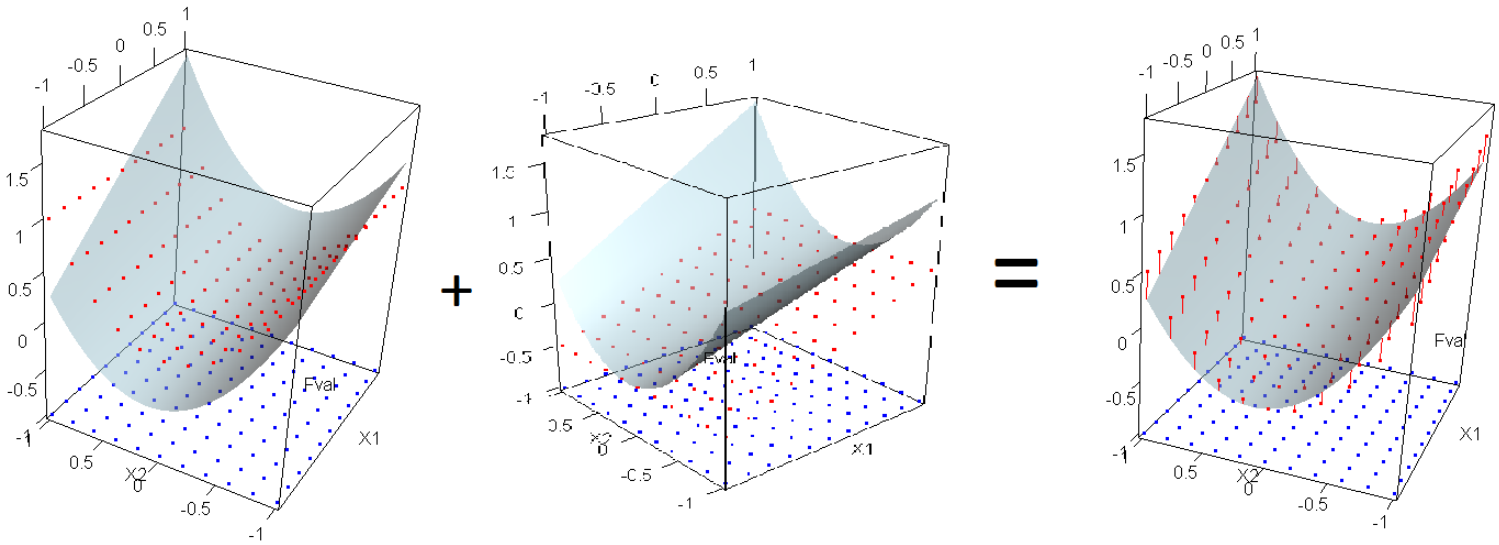
Testing for Additivity

- We can also test for various forms of additivity

Total Additivity: $H_0 : F(\boldsymbol{X}_1, \boldsymbol{X}_2) = F_1(\boldsymbol{X}_1) + F_2(\boldsymbol{X}_2)$

Partial Additivity: $H_0 : F(\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3) = F_1(\boldsymbol{X}_1, \boldsymbol{X}_3) + F_2(\boldsymbol{X}_2, \boldsymbol{X}_3)$

- Define the test set as a grid of test points, viewed as factor levels in a traditional ANOVA set-up, and derive the test-statistics accordingly



Feature Importance

Feature Importance

- Software to compute random random predictions typically also includes a procedure to produce variable importance scores
  - Another reason for their sustained popularity
- Most commonly, Breiman's original out-of-bag (OoB) importance measure is used:

$$VI_{X_1} = \frac{1}{n}\sum_{i=1}^{n}\left(F_{x^*}(X_1, \ldots, X_p) - F_{x^*}(X_1^{\pi}, X_2, \ldots, X_p)\right)^2$$

- Many known problems, especially with overstating importance of correlated features (Strobl et al. 2007, 2008; Nicodemus et al. (2010); Biau and Scornet (2015) for overview)

OoB Importance with Correlation

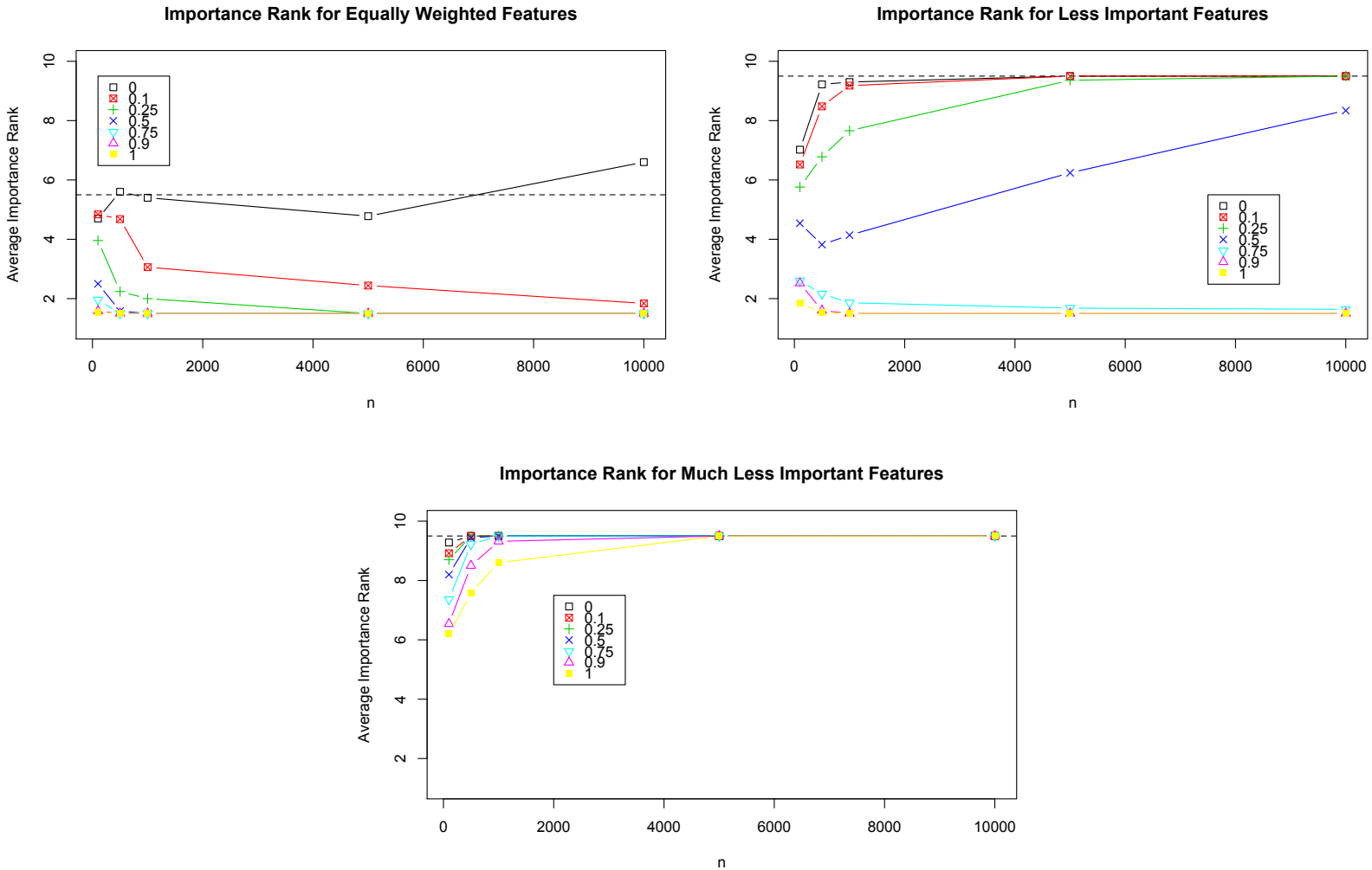


Figure: Average OoB importance of X_9 and X_{10} in additive linear model $y_i = \sum_{i=1}^{10} \beta_i x_i + \epsilon_i$ with $\beta_1 = \cdots = \beta_8 = 1$, $\beta_9 = \beta_{10} = 1, 0.8$, and 0.5, resp.

OoB Importance with Correlation

- Preference for correlated features particularly problematic given the backward-elimination-type procedures commonly employed in practice
  - Might instead prefer something akin to *t* and *F-tests* in linear models: correlated features show weaker marginal importance but significant joint importance
- Proposal: Instead of building, followed by permuting and predicting, permute (or *hold-out*) and rebuild to measure impact on model
  - Akin to extending random feature eligibility to entire trees instead of individual splits

Hold-out Forests

- Structured Hold-out (SHO): Given ensemble of size m, build $\approx \frac{m}{d+1}$ trees excluding (or permuting) each feature and one set with all features
- Random Hold-out (RHO): For each tree in the ensemble, include feature X_i with probability p_i.
- Can define different importance measures:

$$VI_{pred}(X_i) = \frac{1}{n_T} \sum_{j=1}^{n_T} \left(RF_{-i}(x_{T_j}) - RF(x_{T_j}) \right)^2$$
$$VI_{MSE}(X_i) = \frac{1}{n_T} \sum_{j=1}^{n_T} \left(SE_{-i}(x_{T_j}) - SE(x_{T_j}) \right)^2$$

  where $SE_{-i}(x_{T_j}) = \left(RF_{-i}(x_{T_j}) - y_{T_j} \right)^2$
  - Fits within previous framework and allows for importance *intervals* instead of only point estimates
  - Also eliminates the "importance stability issue" – commonly misunderstood

Bike Sharing Dataset Example

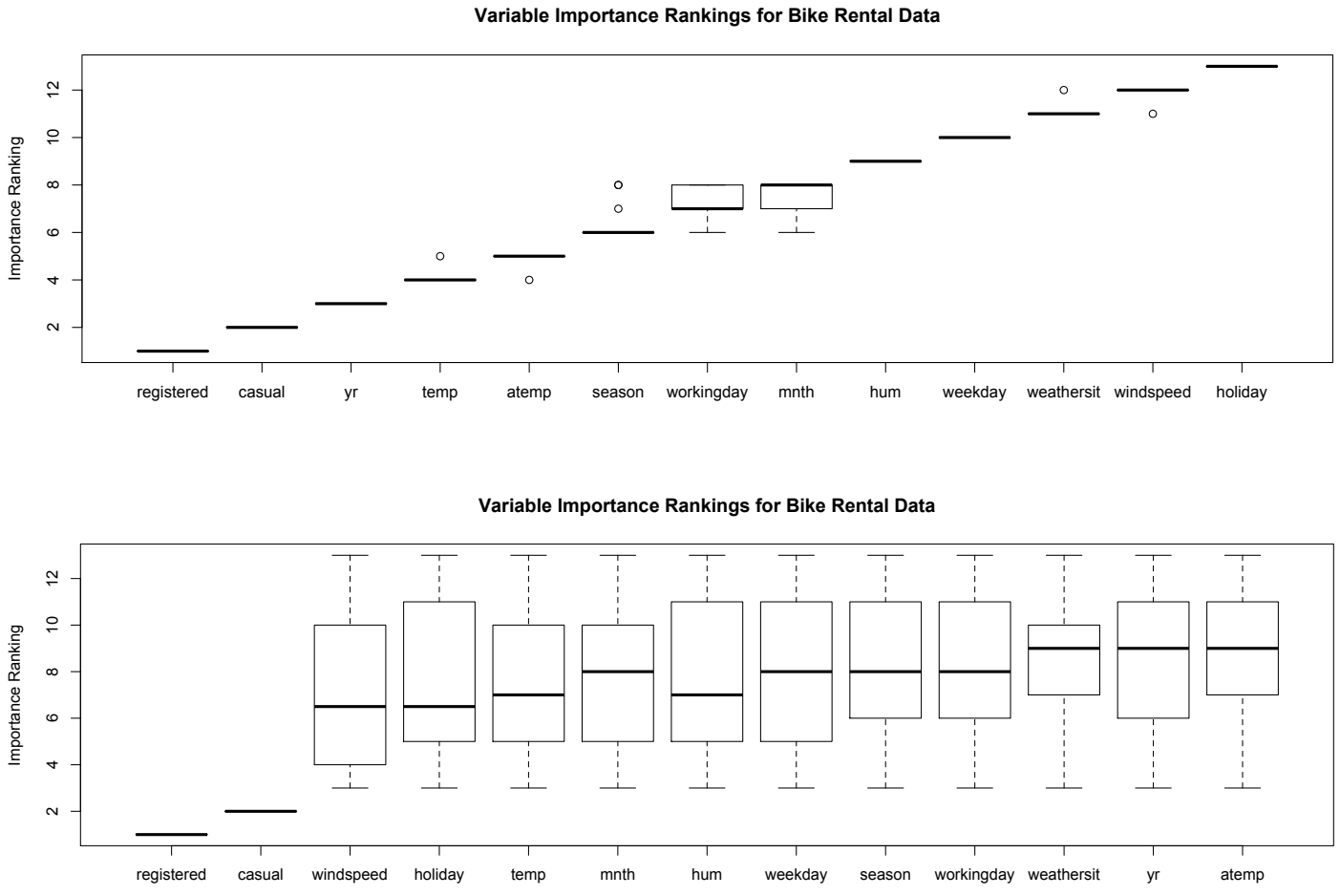


Figure: VI rankings by OoB (above) and SHO (below).

Applications

Application: The eBird Project

Citizen science initiative at Cornell's Laboratory of Ornithology

- Birdwatchers (birders) record observations (time, place, effort, species seen)
- GPS/Time allows addition of Geographic Information Systems (GIS) and weather covariates.
- Currently n > 100 million records and p > 1000.
- Goals:
  - Data-driven maps of bird habitat range/migration
  - Hypothesis generation about bird ecology
  - Evidence of changing ecological factors/behavior
  - Forecasts of bird migration

Indigo Bunting Migration

Indigo Bunting Migration

Indigo Bunting Migration

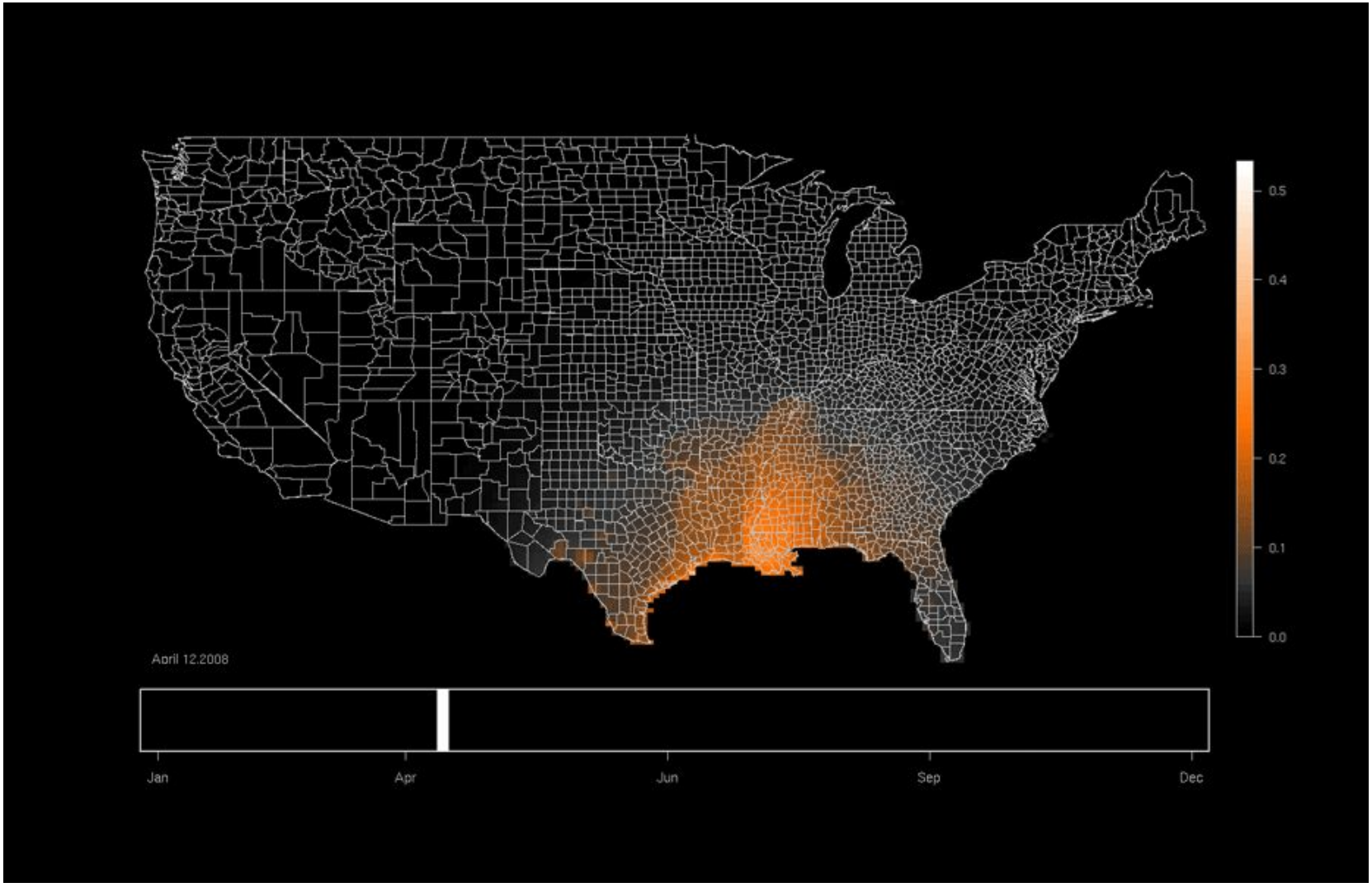
Indigo Bunting Migration

Indigo Bunting Migration

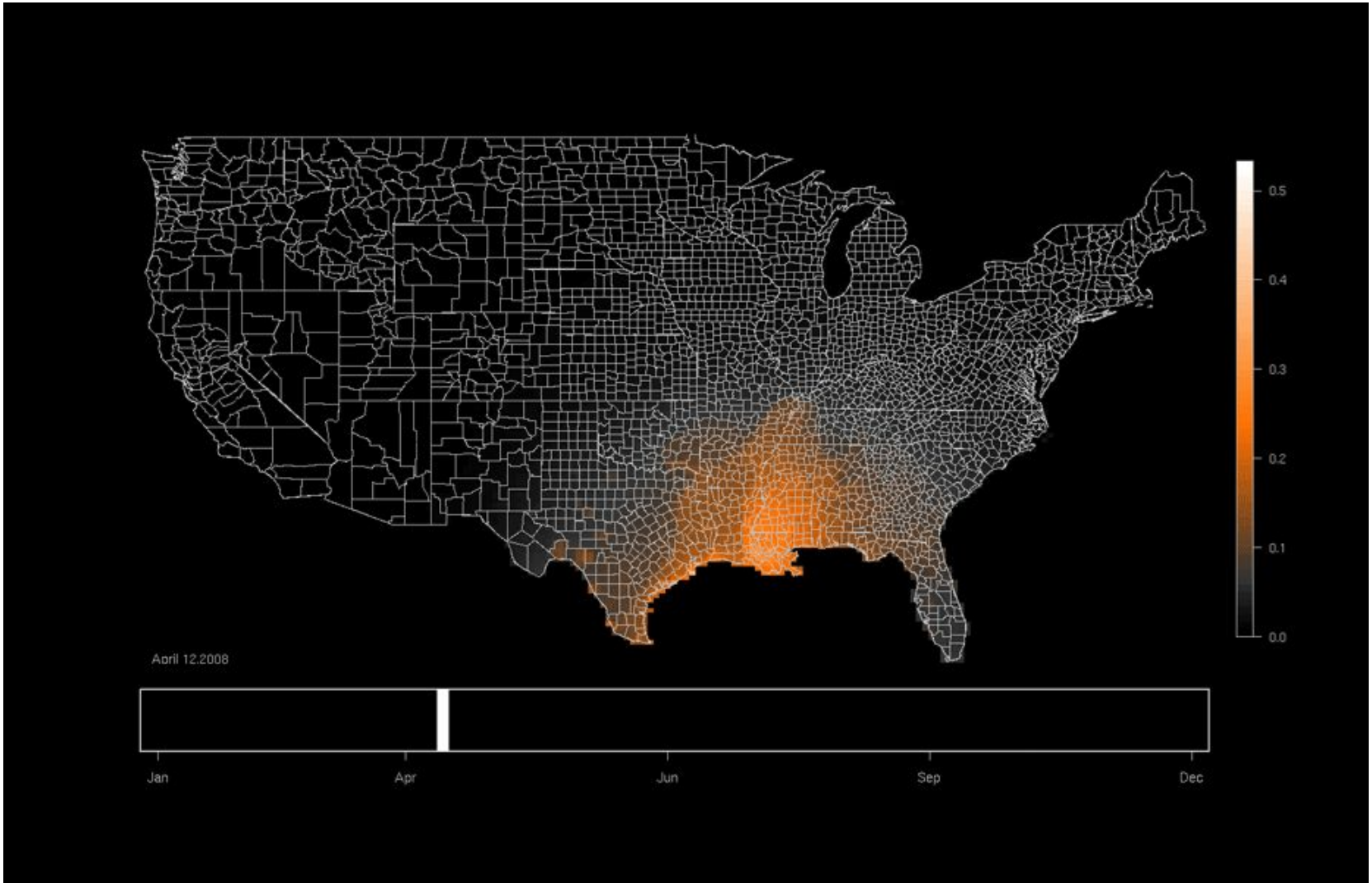
Indigo Bunting Migration

Indigo Bunting Migration

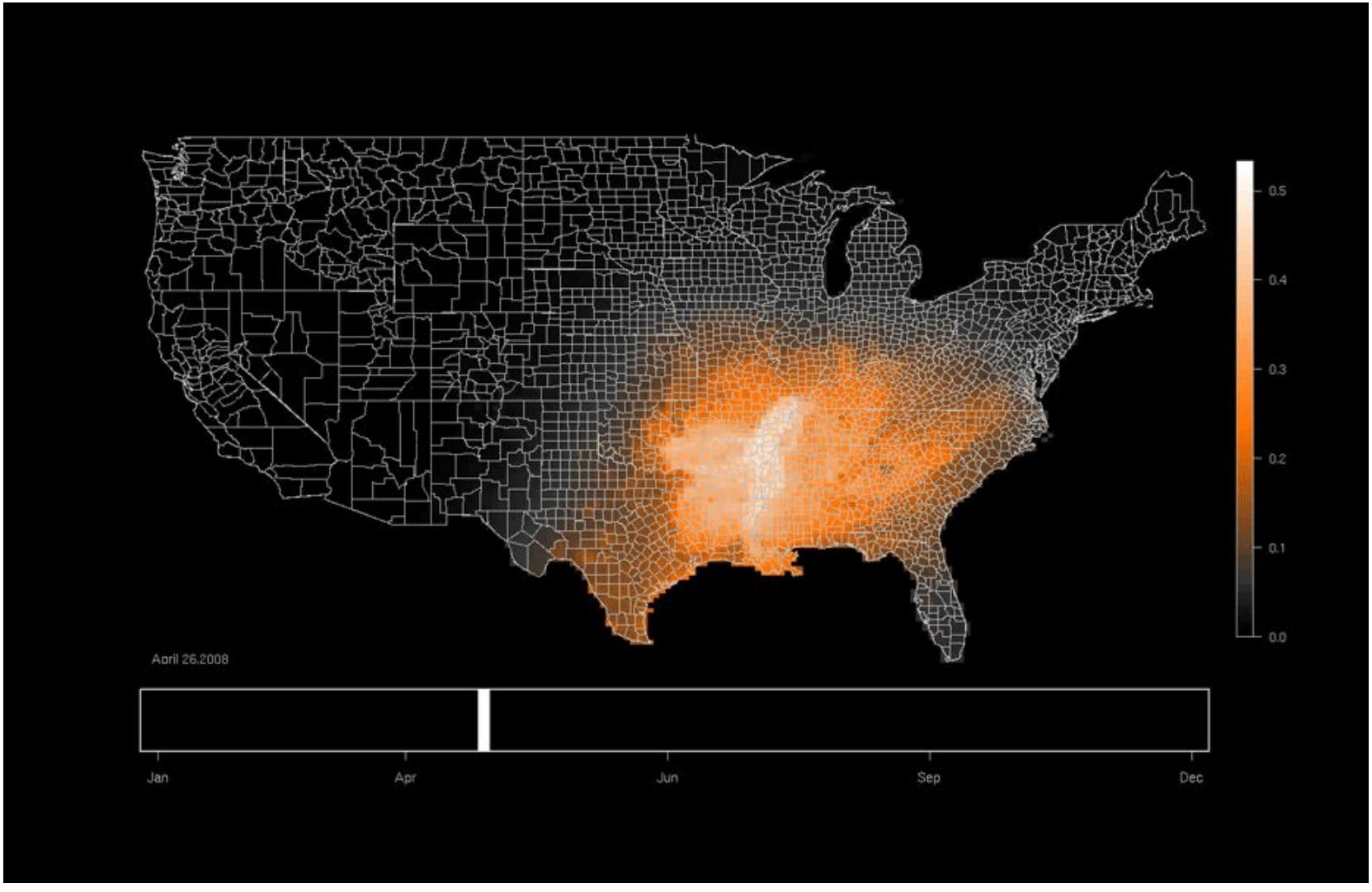
Indigo Bunting Migration



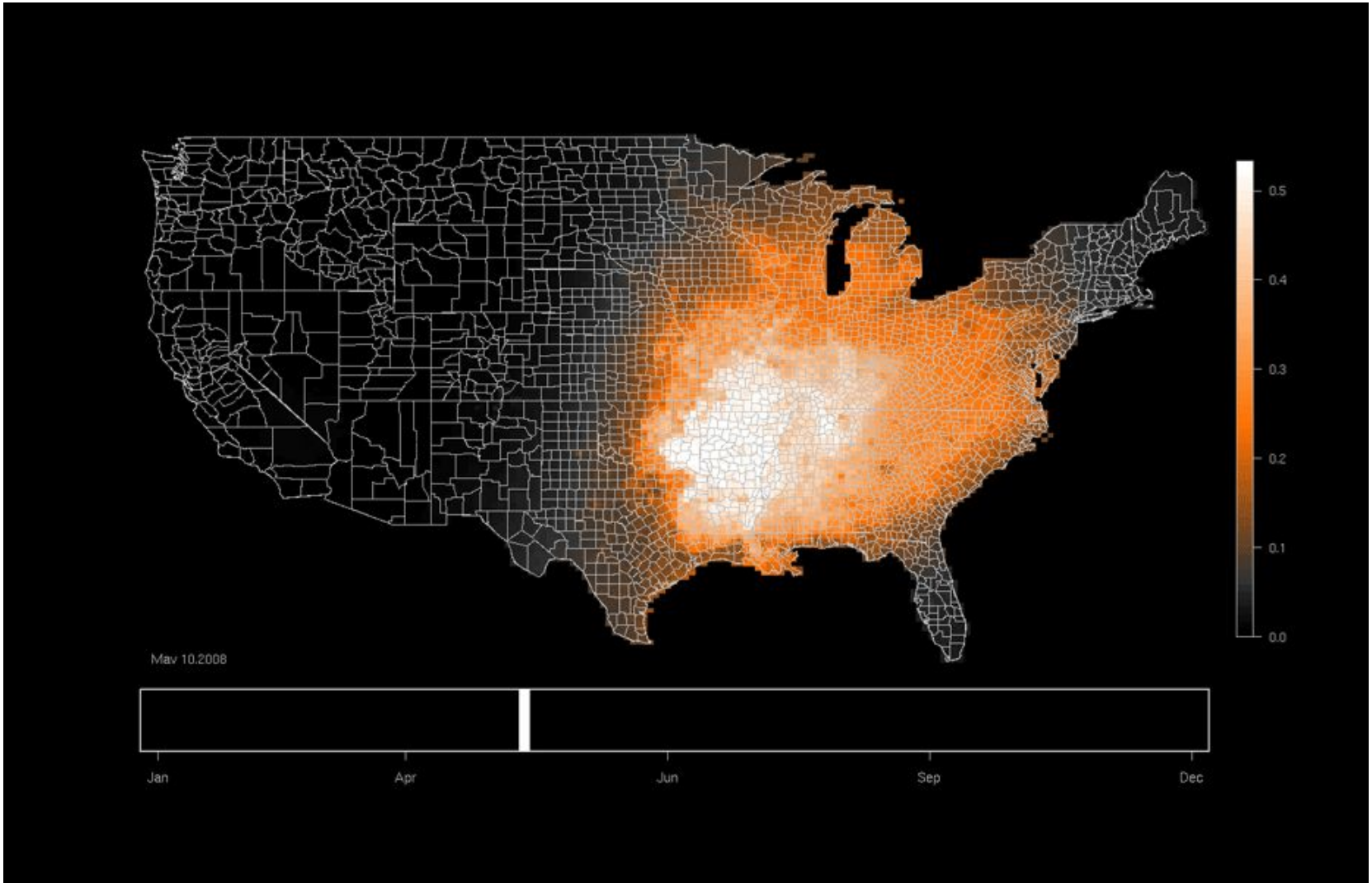
Indigo Bunting Migration



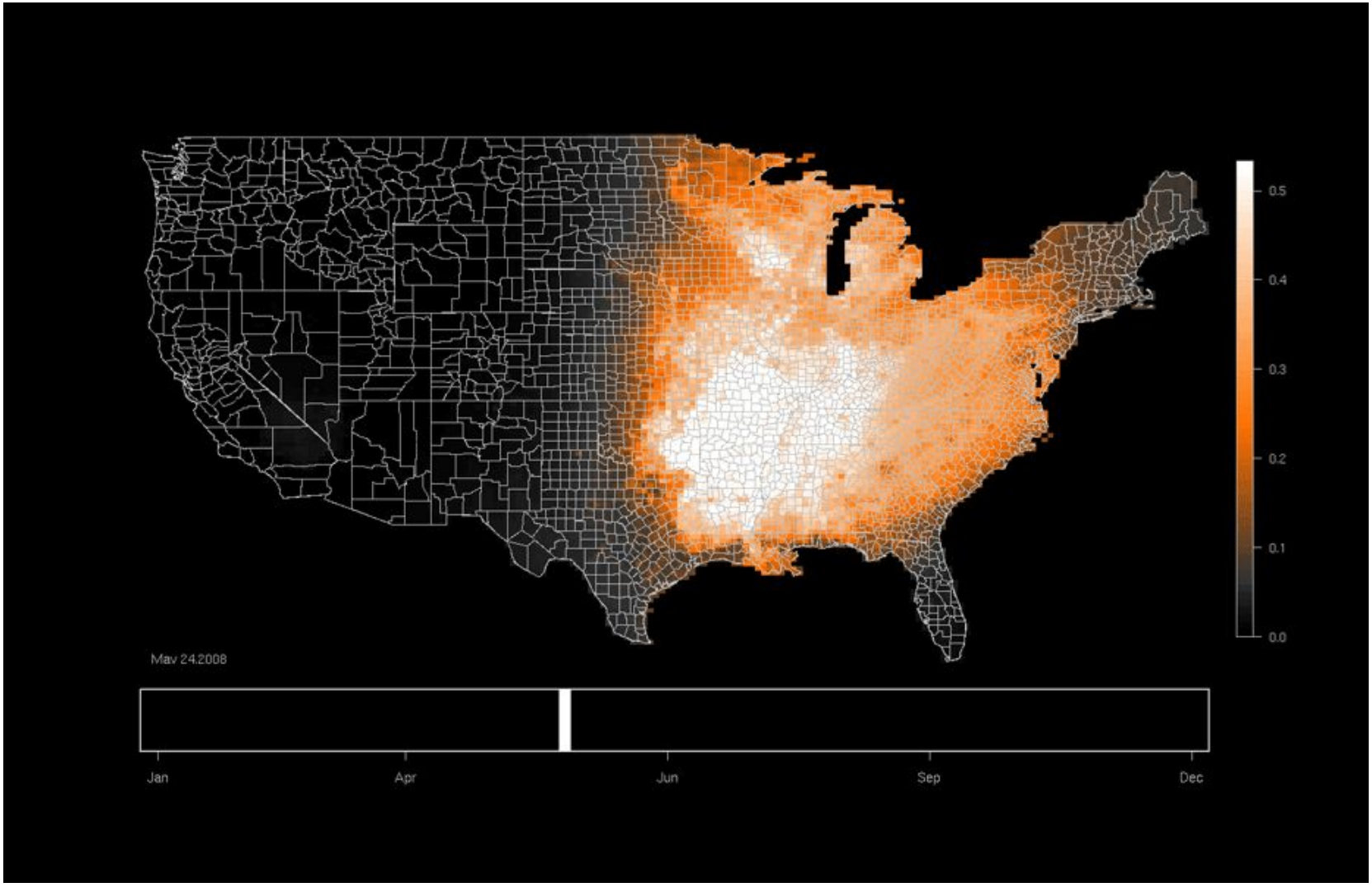
Indigo Bunting Migration



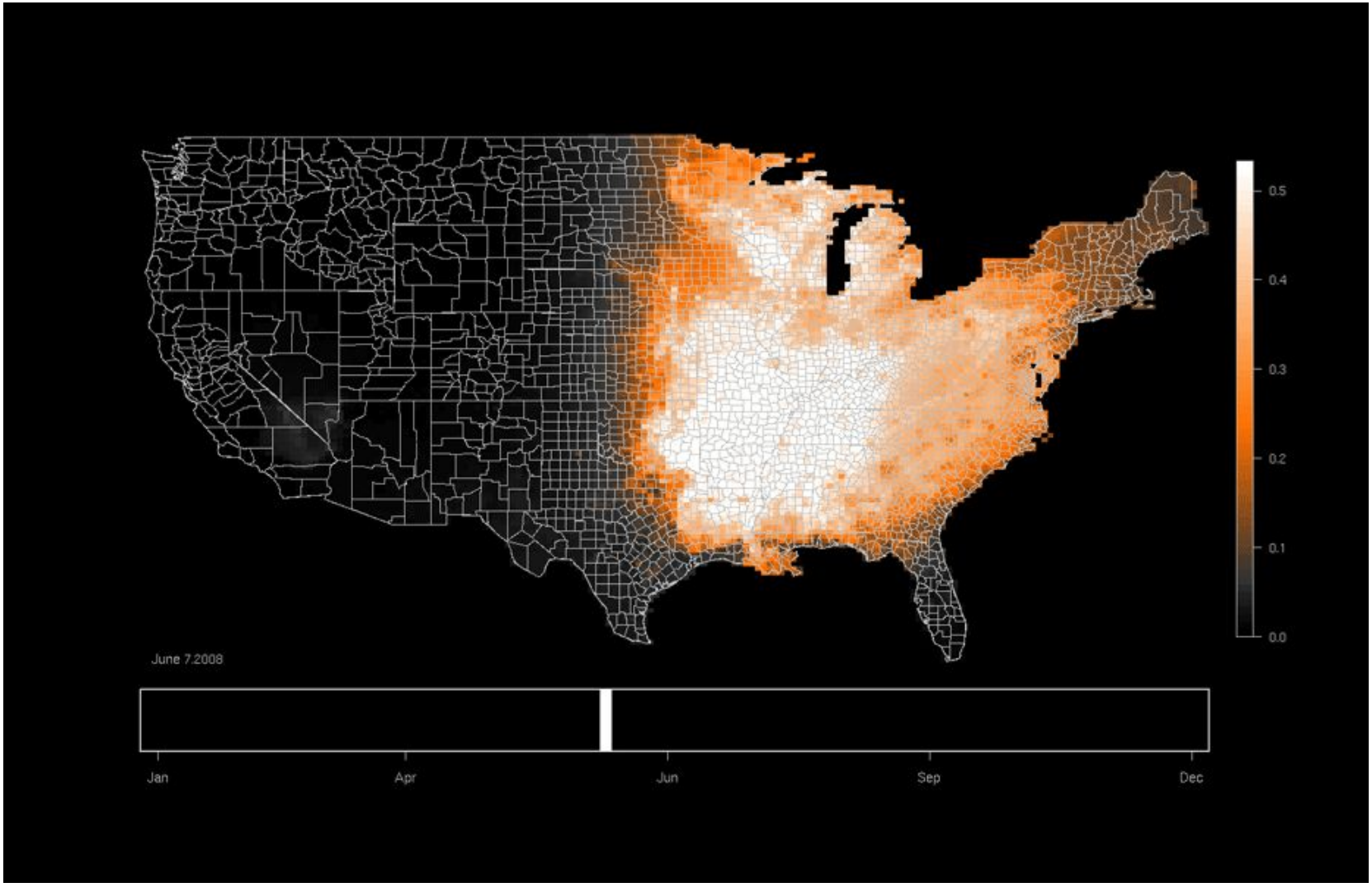
Indigo Bunting Migration



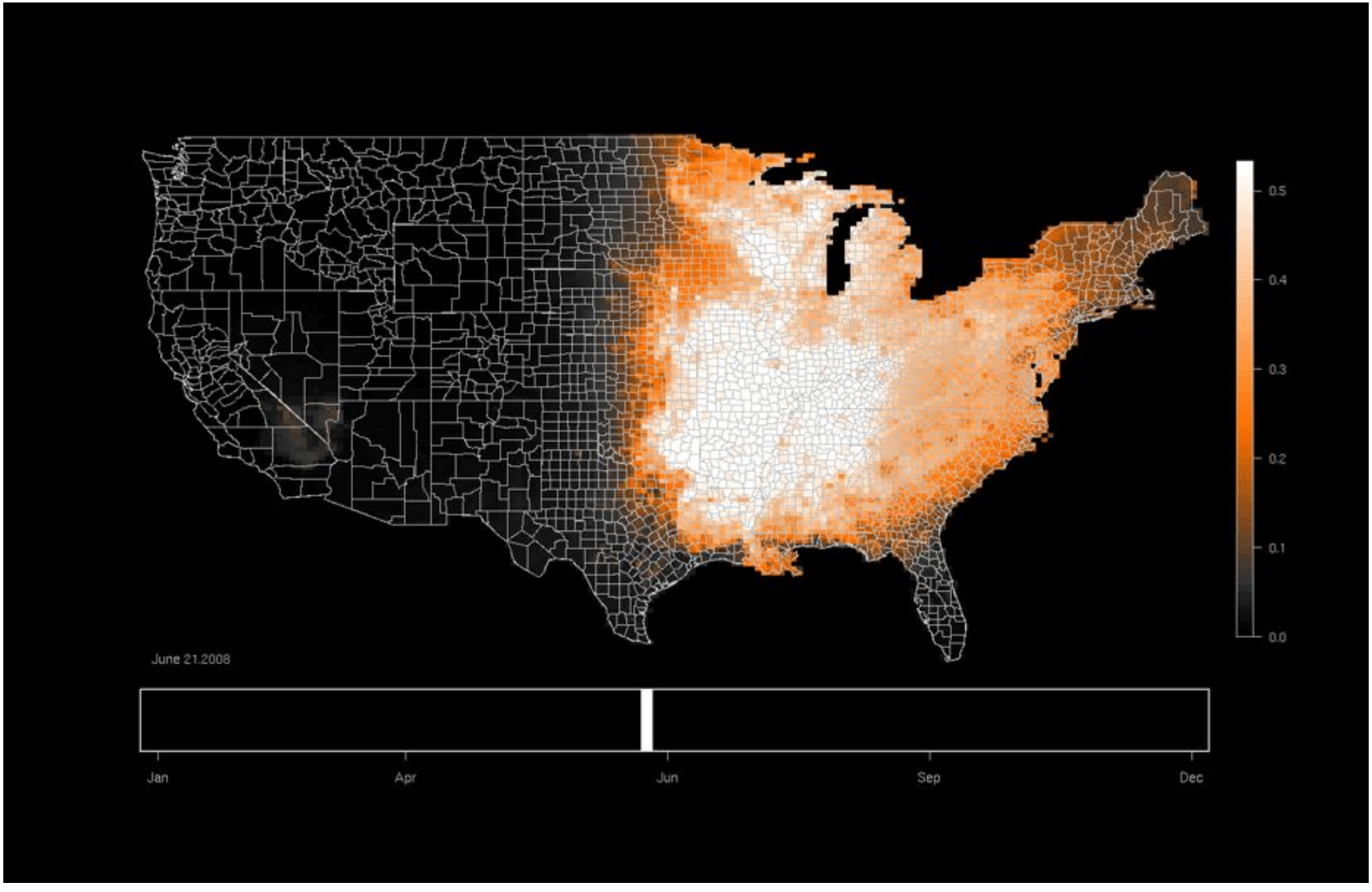
Indigo Bunting Migration



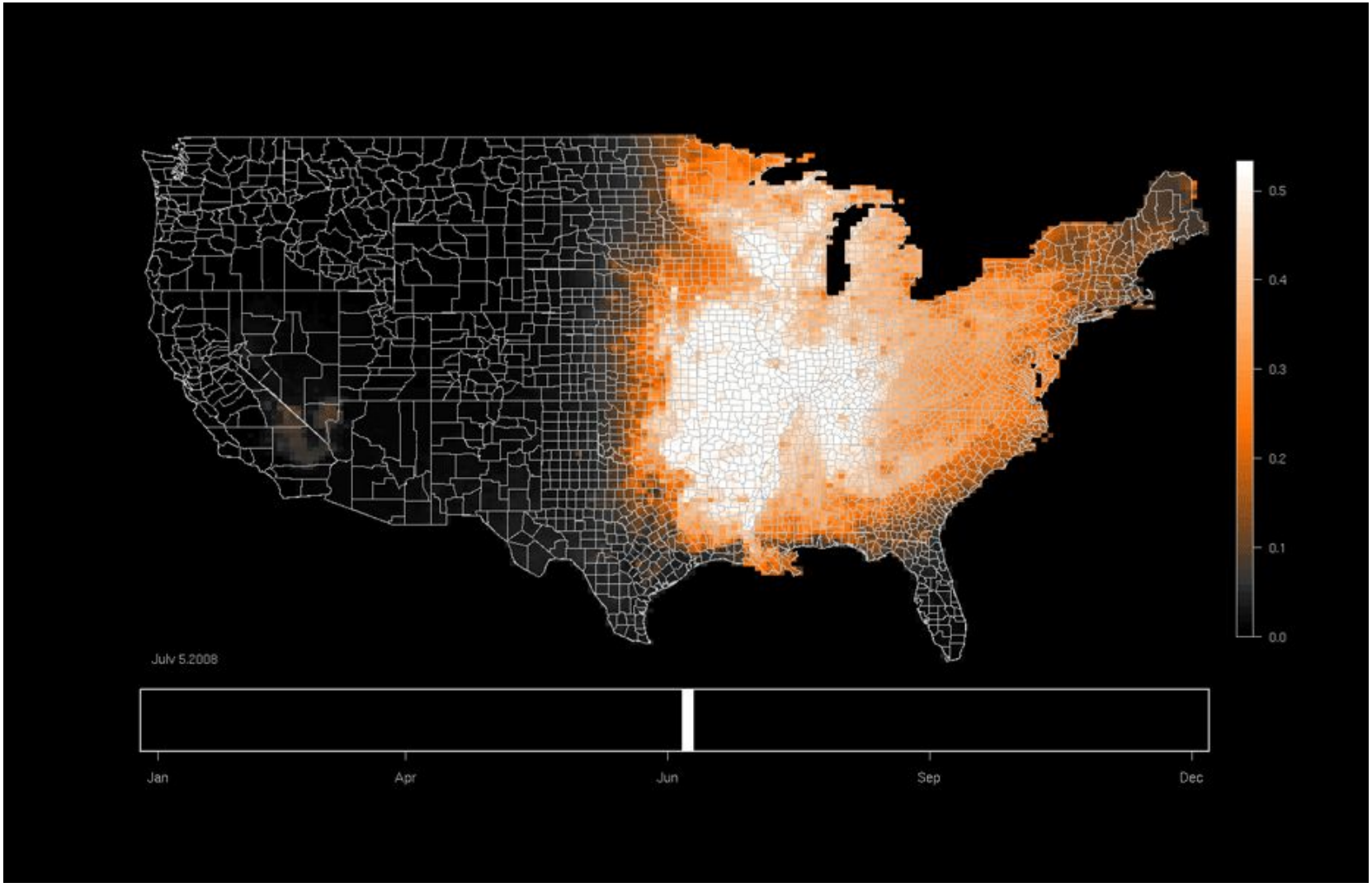
Indigo Bunting Migration



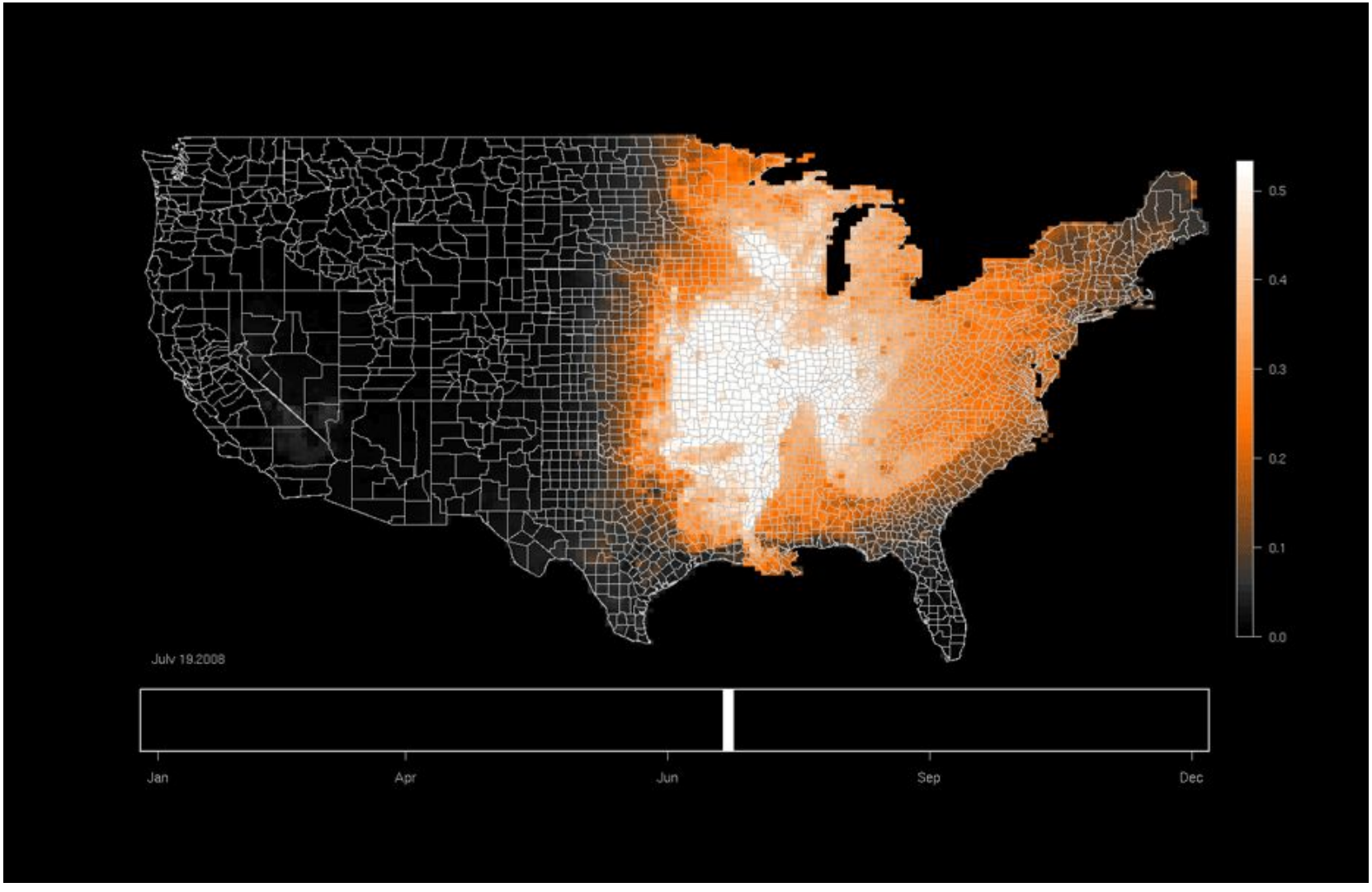
Indigo Bunting Migration



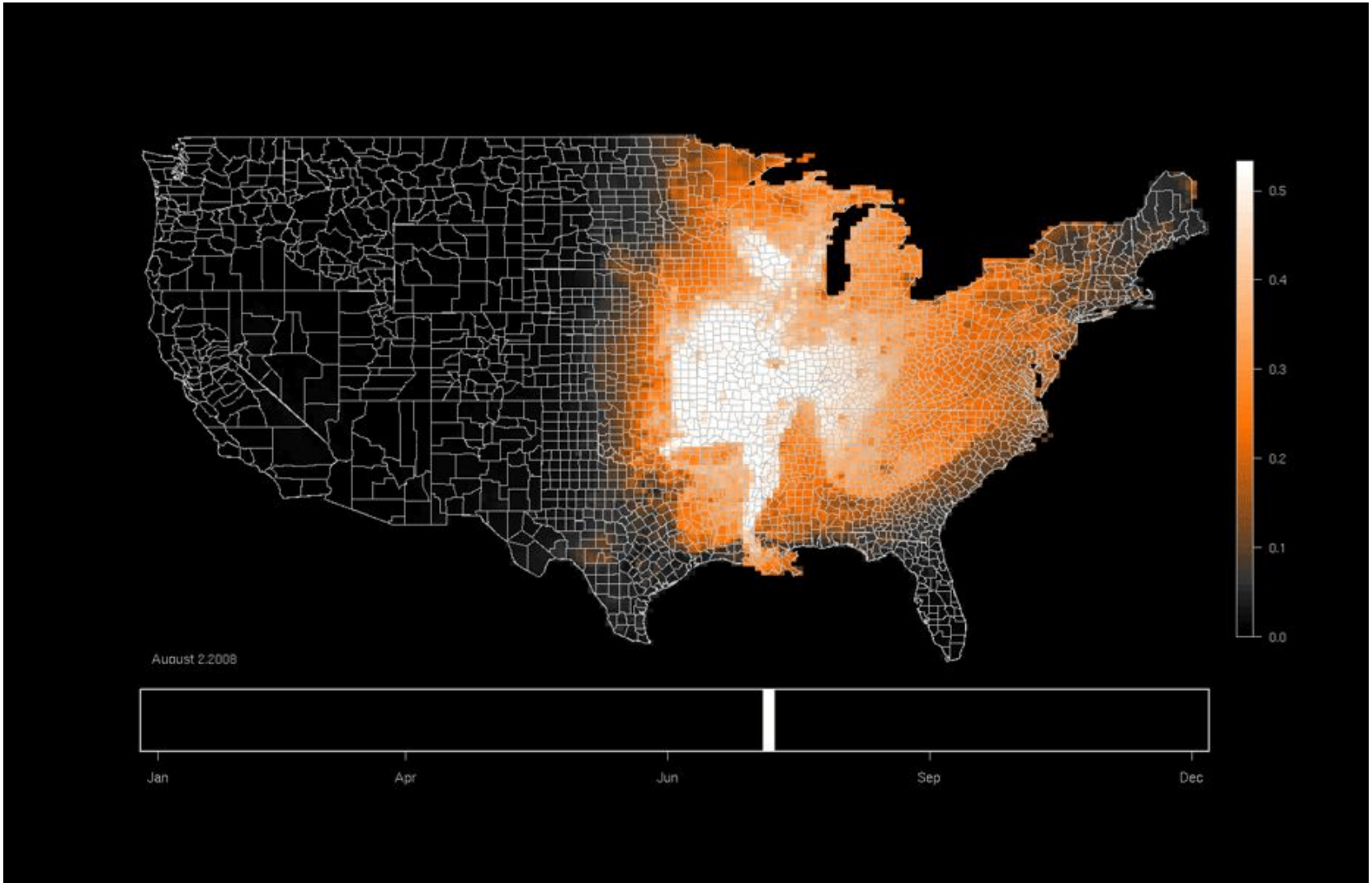
Indigo Bunting Migration



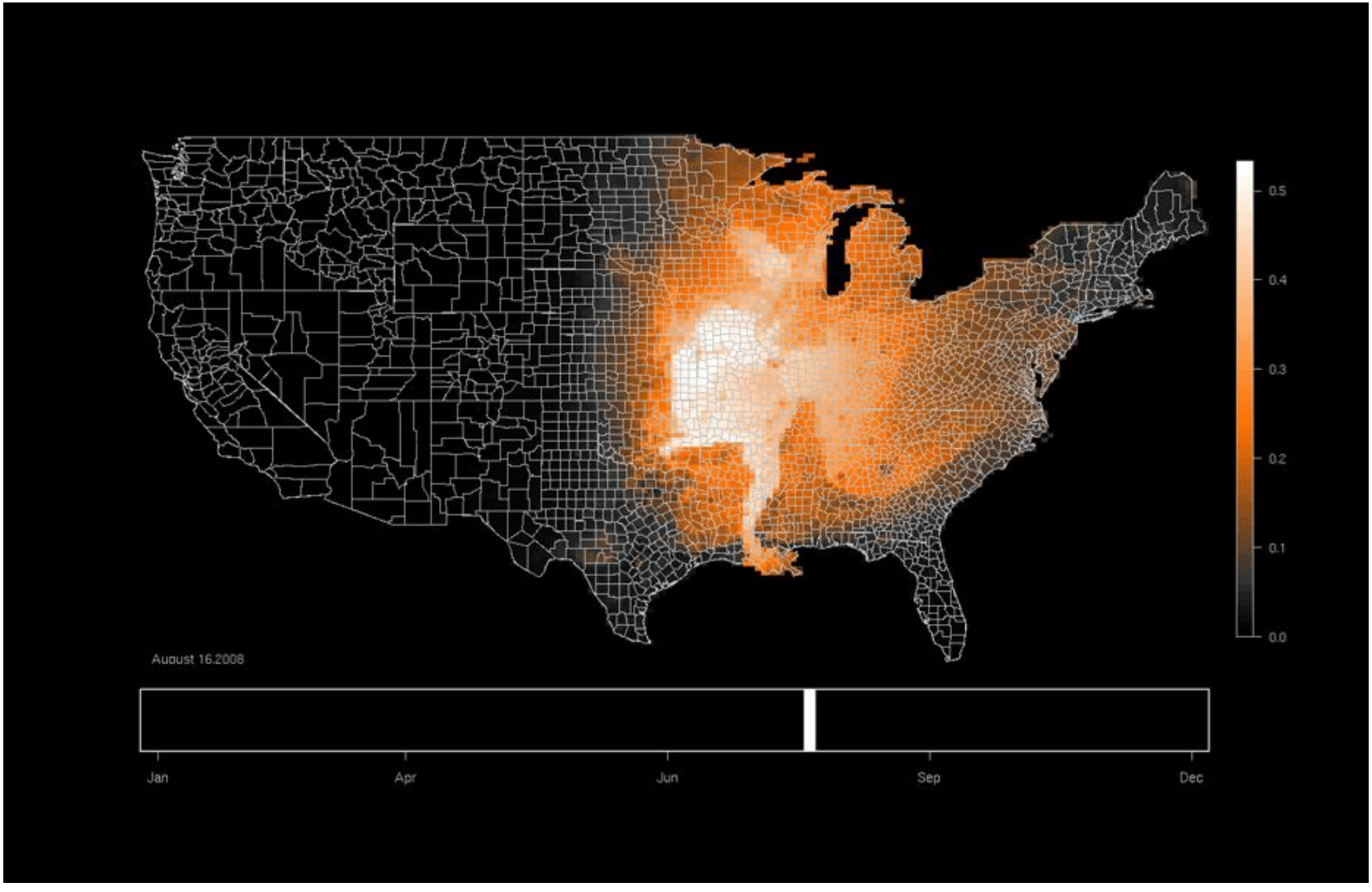
Indigo Bunting Migration



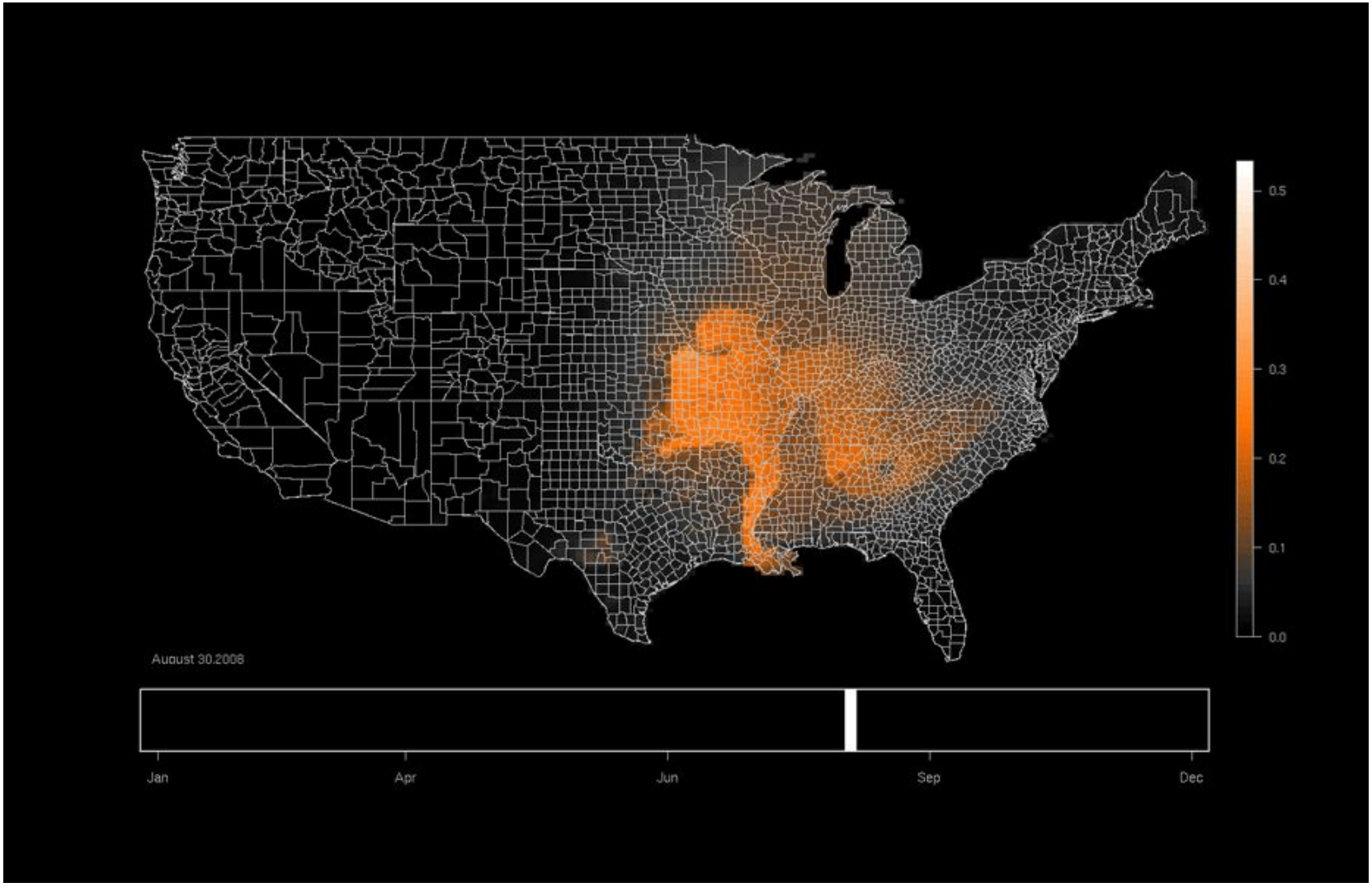
Indigo Bunting Migration



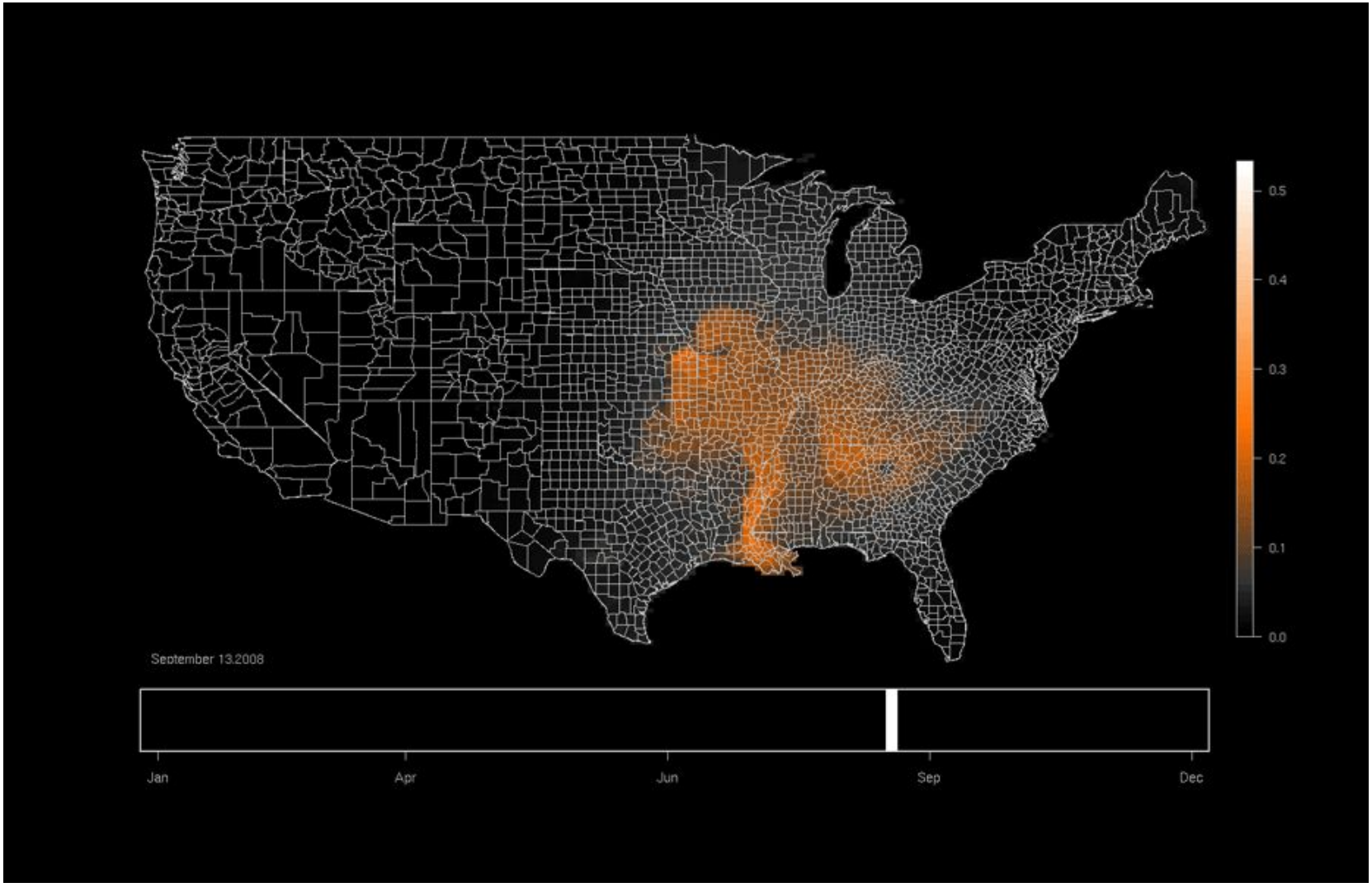
Indigo Bunting Migration



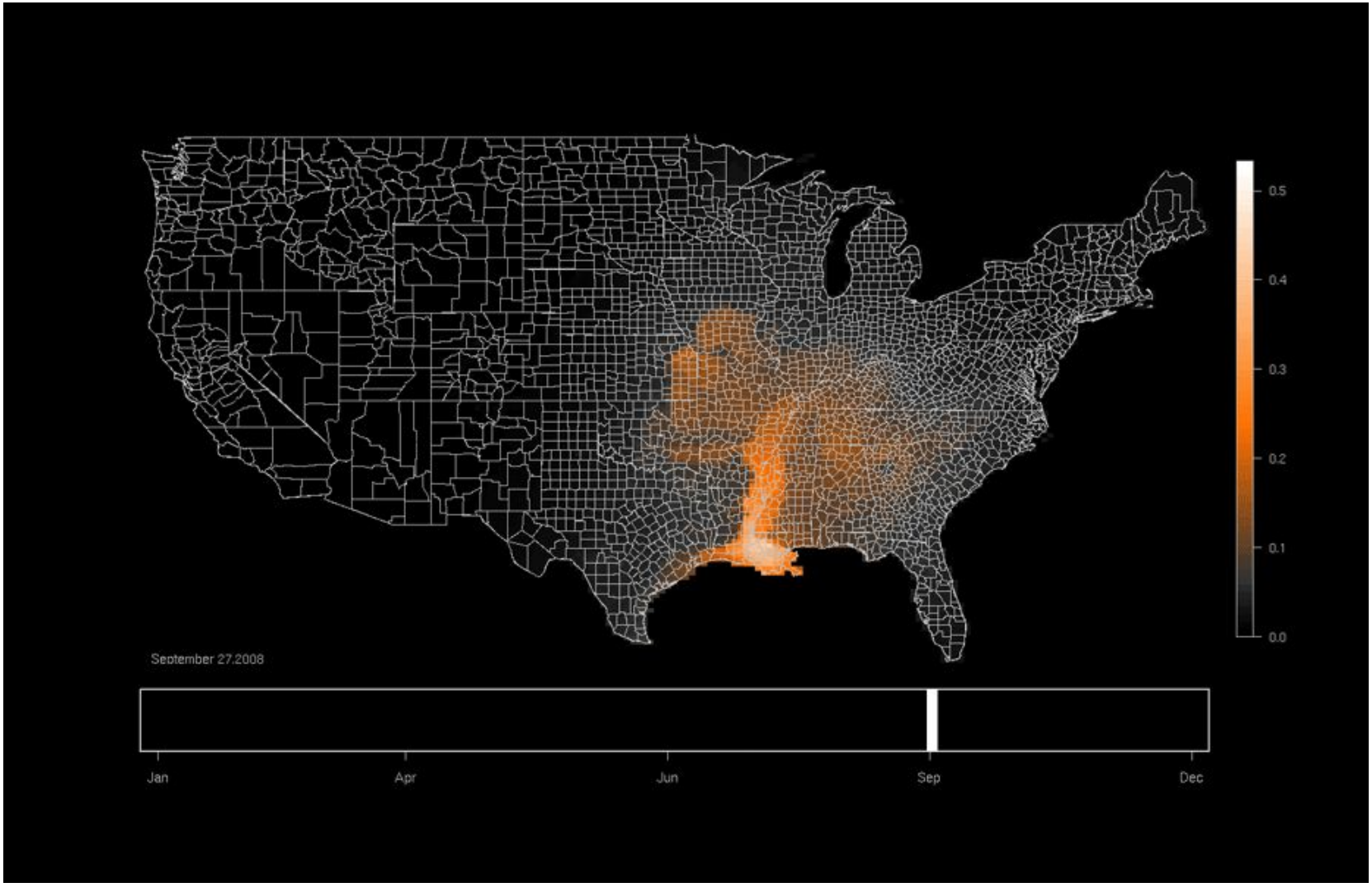
Indigo Bunting Migration



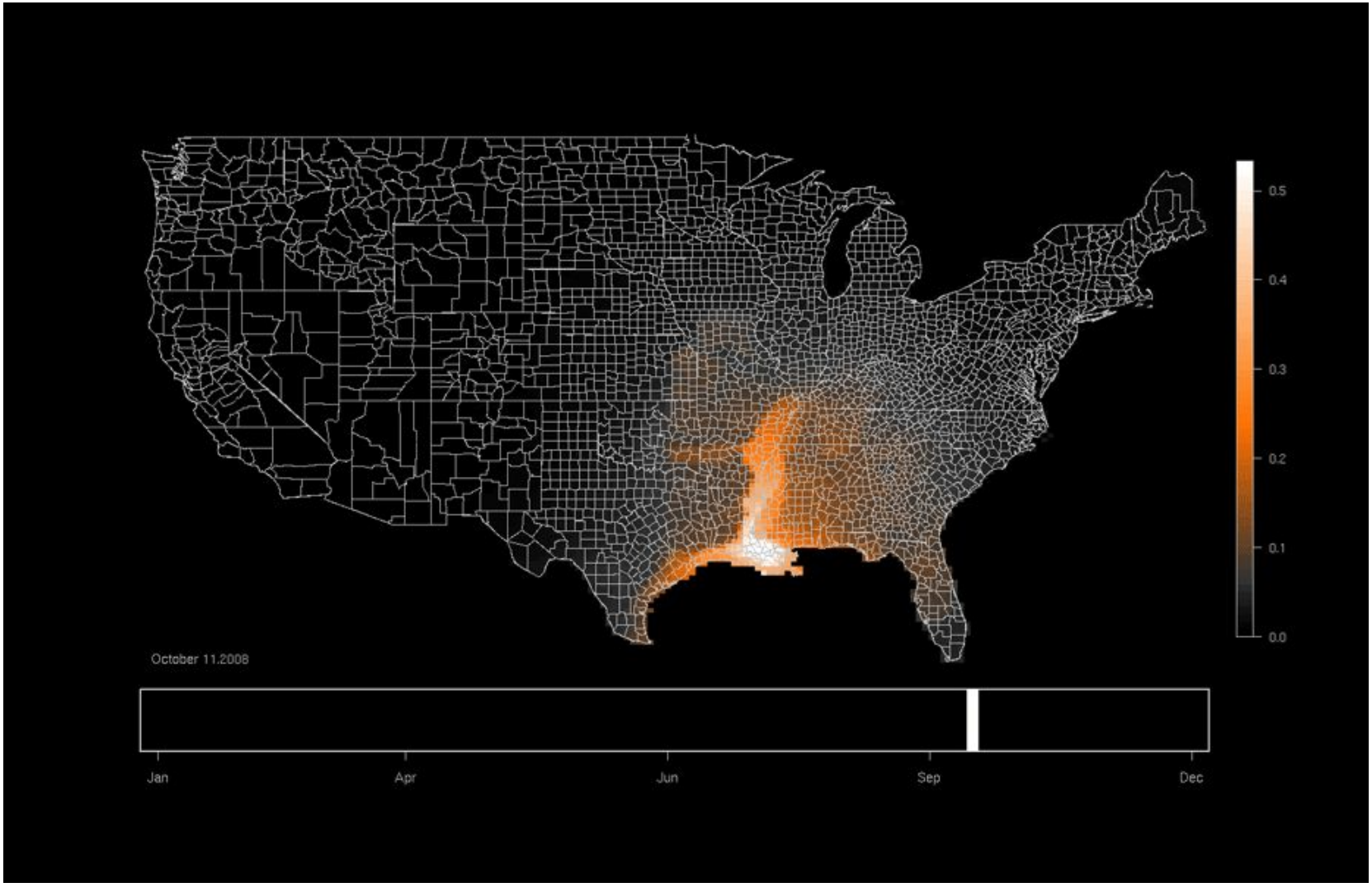
Indigo Bunting Migration



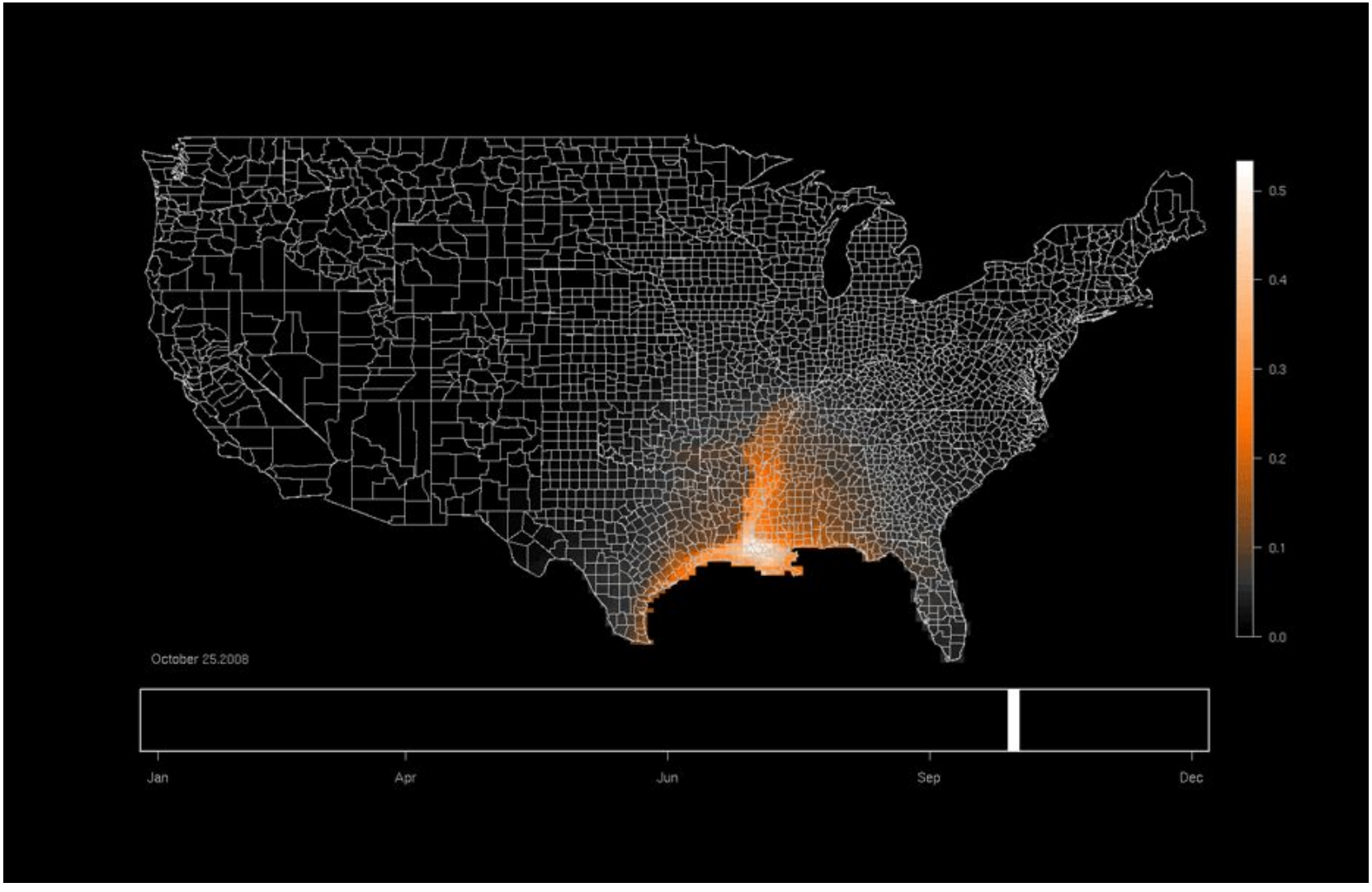
Indigo Bunting Migration



Indigo Bunting Migration



Indigo Bunting Migration



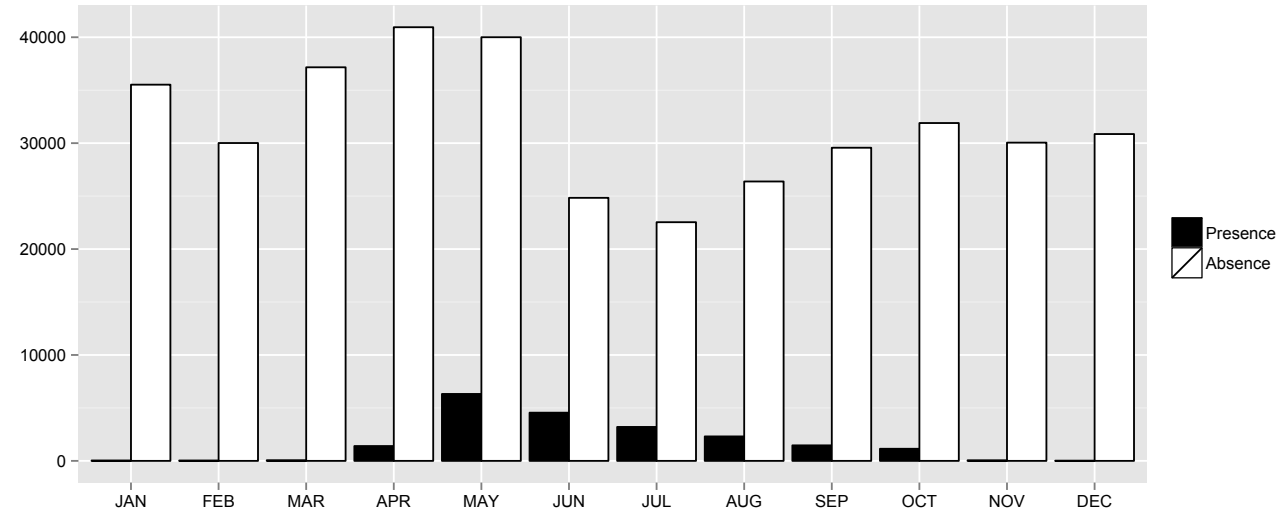
Indigo Bunting Migration

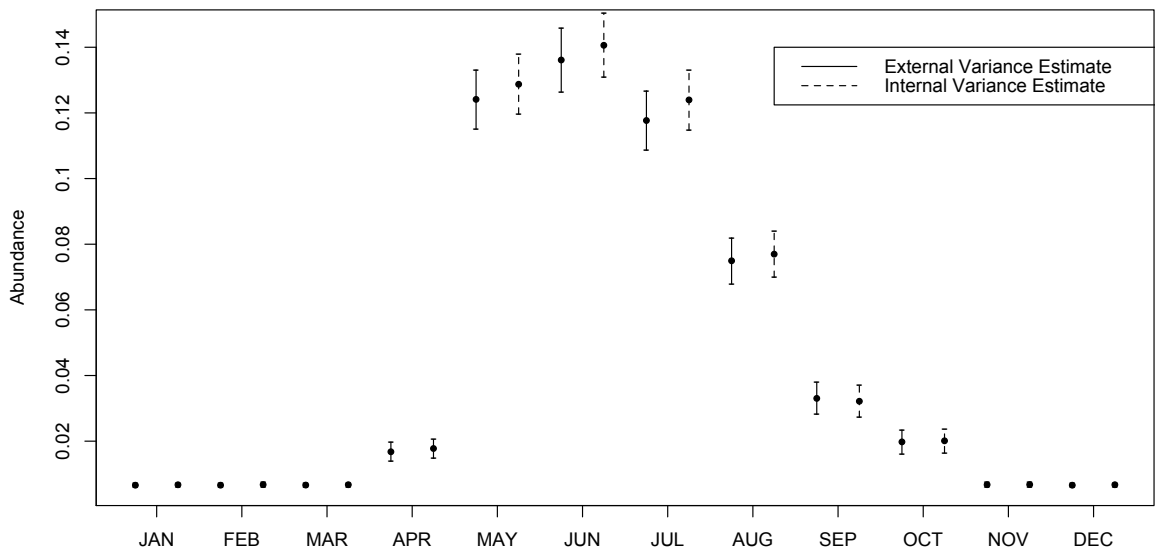
Indigo Bunting Migration

Indigo Bunting Migration

Indigo Bunting Migration

Indigo Bunting Presence/Absence 2010

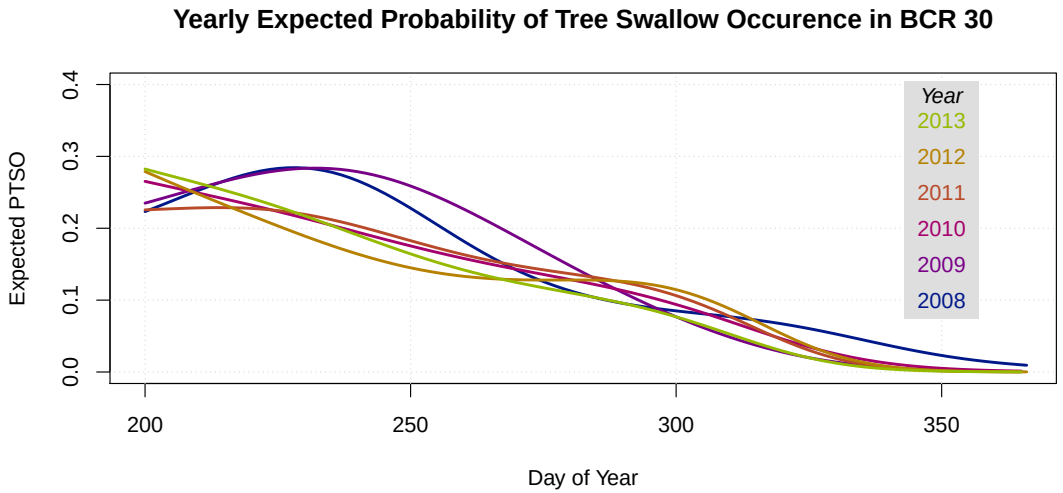




Ebird-Related Testing

- Testing the importance of `month` during the year 2010. (`Time of Year` should be important for migratory species)
  - Month shows up as highly significant, but random values of month can also appear (more slightly) significant
  - Permuting additional variables appears more robust than simple deletion (test statistic values cut in half)
- More recently: looking at effects of Tree Swallow early departure during fall migration along east coast of U.S.
  - Anecdotal evidence that Tree Swallows “left early" in certain years (2008, 2009), though no evidence for why is immediately obvious

Tree Swallow Early Departure



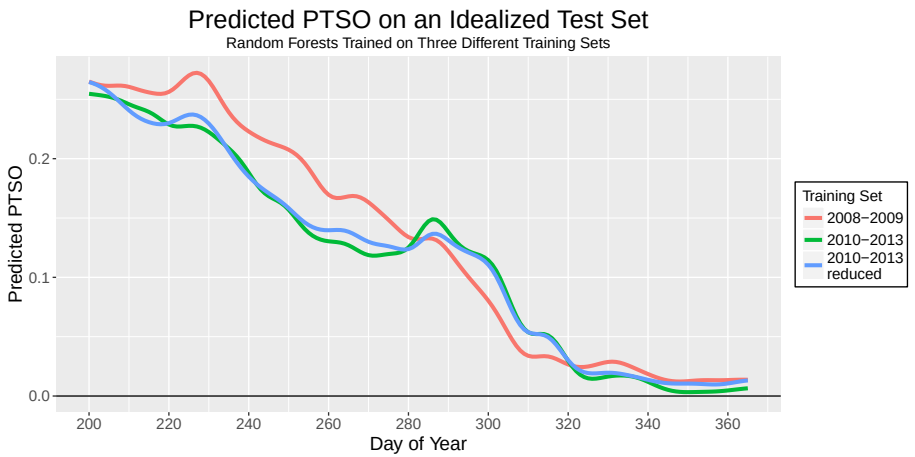


Figure: GAM (top) and RF (bottom) predictions of Tree Swallow occurrence 2008-2013.

Tree Swallow Early Departure

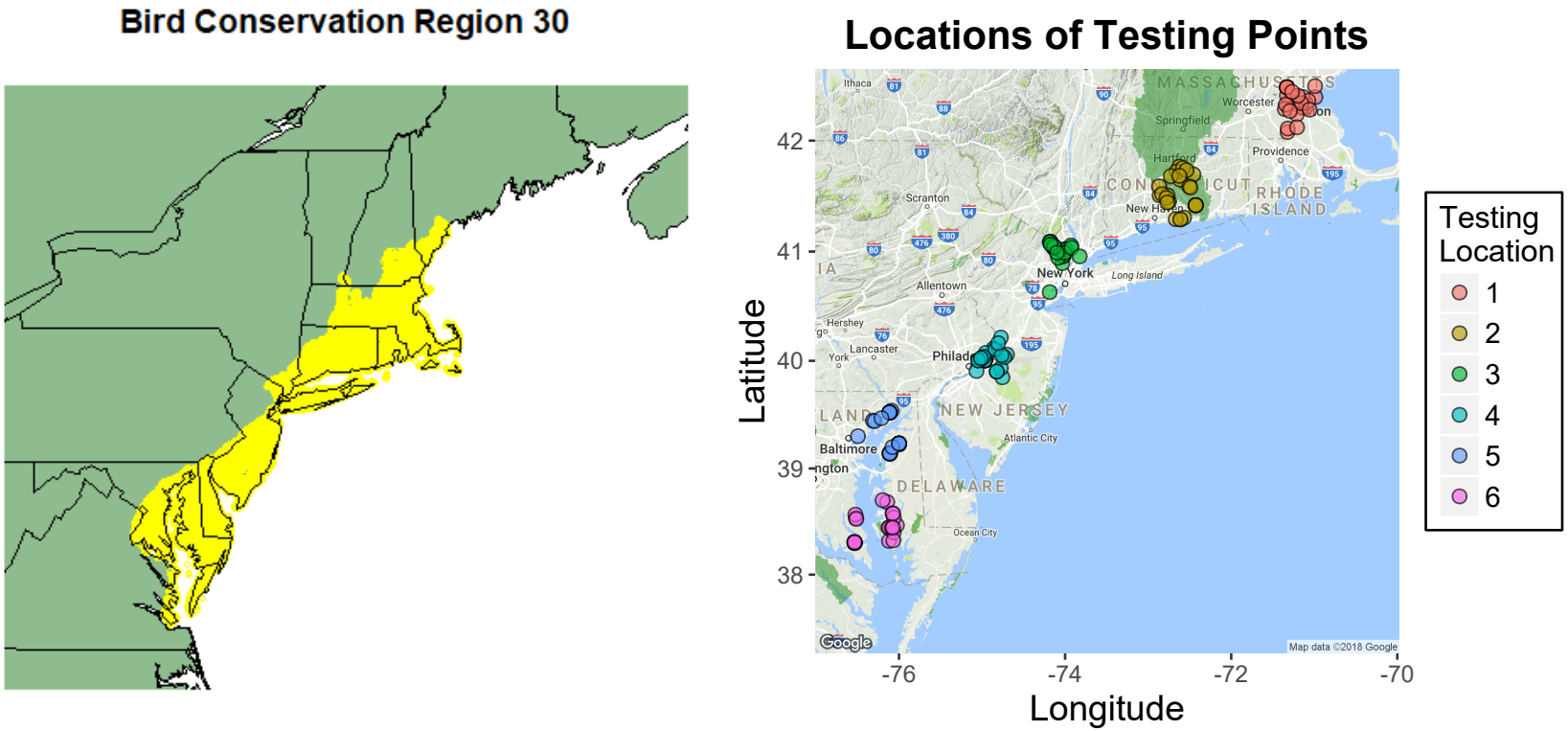


Figure: BCR 30 (left) and specific locations of localized significance tests for `max_temp` (right).

Results

- To test for an effect of year (i.e. were the migration patters in 2008 & 2009 really substantially different from what we would typically expect), we develop a permutation-style test treating the partial effect curves as functional data
  - Substantial differences seen when leaving-in/holding-out `max_temp` vs `day_of_year`, though dropping/permuting `max_temp` not as big an effect as we might want to see
- To de-correlate `max_temp` and `day_of_year`, we test `max_temp_anomaly` at the test points from previous slide and see significance in 5 out of 6 areas

Food for Thought & Future Work

- Testing procedures work well, but can be sensitive to subsample size & accuracy of the estimated covariance parameters
  - Still requires constructing many, many trees
  - Potential alternative methods for estimating those parameters (close relationship to ideas in post-selection inference)
- Currently investigating alternative testing procedures that avoid additional tree construction
- Still appears to be a big gap between theory and practice (e.g. distributions look approximately normal for large subsample sizes; Variance conditions sufficient, but almost certainly not necessary)

SuRFIn R Package

Surfin

Statistical Inference for Random Forests

 Download .tar.gz

 View on GitHub

Description

This R package computes uncertainty for random forest predictions using a fast implementation of random forests in C++. Two variance estimates are provided: U-statistic based (Mentch & Hooker, 2016) and infinitesimal jackknife (Wager, Hastie, Efron, 2014), the latter as a wrapper to the authors' R code randomForestCI.

Check out a demo: How Uncertain Are Your Random Forest Predictions?

Authors and Contributors

Sarah Tan @shftan, David Miller @d-miller, Giles Hooker @gileshooker, Lucas Mentch @LMentch

Relevant Work

• Mentch, Lucas, and Giles Hooker. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *Journal of Machine Learning Research*, 17(26), 1-41.

• Mentch, Lucas, and Giles Hooker. "Formal hypothesis tests for additive structure in random forests." *Journal of Computational and Graphical Statistics* (2017): 1-9.

• Mentch, Lucas and Giles Hooker. "Hold-out Forests for Consistent Feature Importance in Random Forests." *In Progress*.

• Coleman, Tim, Lucas Mentch, Dan Fink, Giles Hooker, et al. "Statistical Inference on Tree Swallow Migration with Random Forests" *In Progress*.

Contact Info

Lucas Mentch
Assistant Professor
Department of Statistics
University of Pittsburgh

lkm31@pitt.edu
lucasmentch@gmail.com
lucasmentch.com